مجله علوم دانشگاه تهران
جلد سی‌ام (۱۳۸۳) شماره ۱ (۴۹-۳۳)

مقایسهٔ مدل‌های حاشیه‌ای برای تحلیل داده‌های طولی با پاسخ‌های دودویی

مجتبی گنجعلی[۱]، امید امدادی‌فر[۲]
دانشگاه شهید بهشتی، دانشکده علوم ریاضی، گروه آمار
(دریافت: ۸۱/۴/۹ ؛ پذیرش: ۸۱/۸/۱۴)

چکیده

مدل حاشیه‌ای (Marginal model) یکی از رهیافت‌هائی است که برای تحلیل داده‌های طولی (Longitudinal data) بکار می‌رود. در این مدل، همبستگی موجود بین پاسخ‌ها برای هر آزمودنی (Individual) به عنوان پارامتر درنظر گرفته می‌شود تا تحلیل‌های معتبری را نتیجه دهد. مدل‌های حاشیه‌ای مختلفی که برای تحلیل داده‌های طولی با پاسخ‌های دودویی ذکر شده‌اند ازجمله مدل‌های حاشیه‌ای با لگ نسبت بخت حاشیه‌ای (Marginal odds ratio)، لگ نسبت بخت شرطی (Conditional odds ratio)، لگ نسبت وابستگی (Dependence ratio)، مدل پروبیت چند متغیره (Multivariate probit model) و روش معادلات برآوردگر تعمیم یافته [Generalized Estimating Equations (GEE)] مرور و مقایسه شده‌اند. مانده‌هایی برای بررسی نیکویی برازش این مدل‌ها معرفی گردیده‌است. همچنین مدل‌های ذکرشده را درمثالی کاربردی برازش داده‌ایم.

واژه‌های کلیدی: *مطالعات طولی، مدل‌های حاشیه‌ای، معادلات برآوردگر تعمیم‌یافته، نسبت وابستگی، نسبت بخت حاشیه‌ای، نسبت بخت شرطی، مدل پروبیت چند متغیره*

۱- مقدمه

داده‌های طولی داده‌هایی هستند که درآنها پاسخ‌ها برای هرآزمودنی درطول زمان تکرار شده‌اند و چون اغلب بین پاسخ‌ها برای هر آزمودنی خاص همبستگی وجود دارد نیاز به روش‌های ویژه‌ای برای تحلیل وجود دارد تا بتوان استنباط‌های معتبری به‌دست آورد. چندین رهیافت برای تحلیل چنین داده‌هایی بکار می‌روند که ازجمله آنها می‌توان مدل‌های اثرات تصادفی (Random effects models)، مدل‌های انتقالی (Transition models) و مدل‌های حاشیه‌ای را نام برد (Diggel, *et al.*, 1994).

دربخش دوم مدل حاشیه‌ای را تعریف کرده، سپس انواع مدل‌های حاشیه‌ای را که برای داده‌های طولی دودویی بکار می‌روند در بخش‌های ۳ تا ۶ مرور کرده‌ایم. در بخش ۷ مقایسه‌ای کلی بین مدل‌های حاشیه‌ای ارائه شده، انجام داده‌ایم. از آنجا که در داده‌های طولی دودویی مانده‌هایی که بتوانند با درنظر گرفتن همبستگی پاسخ‌های یک آزمودنی خاص نیکویی برازش مدل را انجام دهند هنوز معرفی نشده‌است، مانده‌هایی برای مدل‌های ذکر شده در بخش ۸ معرفی کرده‌ایم. سپس مدل‌های بیان شده را در بخش ۹ بر روی داده‌های آسم (Rotnitzky & Wypij, 1994) برازش داده‌ایم.

۲- مدل‌های حاشیه‌ای

مدل‌های حاشیه‌ای مدل‌هایی هستند که در آنها تاثیر متغیرهای تبیینی بر پاسخ‌ها به طور مجزا از همبستگی بین پاسخ‌ها برای یک آزمودنی معین مدل‌بندی می‌شود. دراین نوع مدل‌ها، امید حاشیه‌ای ـ $E(Y_{ij})$ ـ به صورت تابعی از متغیرهای تبیینی مدل‌بندی می‌شود. امید حاشیه‌ای همان پارامتری است که درمطالعه مقطعی نیز مدل‌بندی می‌شود(Agresti, 1990).

دریک مدل حاشیه‌ای امید حاشیه‌ای پاسخ ـ $\mu_{ij} = E(Y_{ij})$ ـ به متغیرهای تبیینی ـ x_{ij} ـ بصورت $h(\mu_{ij}) = x'_{ij}\beta$ وابسته است؛ که در آن h تابع پیوند (Link function) شناخته شده‌ای است. به عنوان مثال h برای پاسخ‌های دودویی، لوجیت می‌باشد. همچنین واریانس حاشیه‌ای به میانگین حاشیه‌ای به صورت $\mathrm{var}(Y_{ij}) = v(\mu_{ij})\phi$ وابسته است که v تابع واریانس شناخته شده و ϕ پارامتر مقیاسی است که ممکن است، علاوه بر دیگر پارامترها نیاز به برآورد داشته باشد و همبستگی Y_{ij} و Y_{ik} به صورت تابعی از میانگین حاشیه‌ای آنها و پارامترهای اضافی ـ α - است یعنی:

$$Corr\,(Y_{ij}, Y_{ik}) = \rho\,(\mu_{ij}, \mu_{ik}; \alpha)$$

که $\rho(.)$ تابع شناخته شده‌ای از μ_i ها است. ضرایب رگرسیونی حاشیه‌ای ـ β ـ دارای تفسیری مشابه ضرایب رگرسیونی در تحلیل مقطعی هستند. می‌توان گفت مدل‌های حاشیه‌ای برای پاسخهای وابسته تعمیمی از مدل‌های خطی تعمیم یافته(Generalized linear models) برای پاسخ‌های مستقل می‌باشند.

۳- رهیافت بهادر (Bahadur)

بهادر (۱۹۶۱) توزیع توام $Y_i = (Y_{i1}, \ldots, Y_{in})$ که در آن Y_{ij} برای $j=1, \ldots, n$ و $i=1, \ldots, m$ پاسخ‌های دودویی هستند را به صورت زیر در نظر می‌گیرد.

$$\prod_{j=1}^{n} \mu_{ij}^{y_{ij}} (1-\mu_{ij})^{1-y_{ij}} \left(1 + \sum_{j<k} \rho_{ijk} e_{ij} e_{ik} + \sum_{j<k<l} \rho_{ijkl} e_{ij} e_{ik} e_{il} + \ldots + \rho_{i12\ldots n} e_{i1} e_{i2} \ldots e_{in} \right)$$

که درآن:

$$\rho_{i12\ldots n} = E(e_{i1} e_{i2} \ldots e_{in}) \text{ و } \ldots, \rho_{ijk} = E(e_{ij} e_{ik}), \quad e_{ij} = \frac{y_{ij} - \mu_{ij}}{\{\mu_{ij}(1-\mu_{ij})\}^{1/2}}$$

بنا براین توزیع توام پاسخ‌ها بوسیله فرضهایی درباره‌۱-n- " ۲ همبستگی‌های حاشیه‌ای

$$\rho_i = (\rho_{i12}, \ldots, \rho_{in-1n}, \ldots, \rho_{i12\ldots n})'$$

مشخص می‌شود.

نمایش بهادر در بالا برای درجه‌های متفاوتی از وابستگی در میان Y_i ها قابل بیان می‌باشد. مثلا ساختار استقلال نتیجه می‌دهد که تمام همبستگی‌های حاشیه‌ای دودویی و مراتب بالاتر صفر است. اما همبستگی‌های حاشیه‌ای (که درمدل بهادر تعدادشان با افزایش n بسیار زیاد می‌شود) به صورت زیر بوسیله احتمال‌های حاشیه‌ای(میانگین‌های حاشیه‌ای که در پاسخ‌های دودویی احتمال‌های پیروزی حاشیه‌ای‌اند) محدود می‌شوند.

$$A \le Corr(Y_{ij}, Y_{ik}) \le B \qquad (1)$$

که درآن:

$$A = \max\left\{ -\left(\frac{\mu_{ij}\mu_{ik}}{(1-\mu_{ij})(1-\mu_{ik})} \right)^{1/2}, -\left(\frac{(1-\mu_{ij})(1-\mu_{ik})}{\mu_{ij}\mu_{ik}} \right)^{1/2} \right\}$$

و

$$B = \min\left\{\left(\frac{\mu_{ij}(1-\mu_{ik})}{(1-\mu_{ij})\mu_{ik}}\right)^{\frac{1}{2}}, \left(\frac{(1-\mu_{ij})\mu_{ik}}{\mu_{ij}(1-\mu_{ik})}\right)^{\frac{1}{2}}\right\}$$

به‌همین دلیل بسیاری‌از محققین، معیارهایی برای همبستگی درنظر می‌گیرند که موجب می‌شود حدود همبستگی به میانگین‌ها مقید نباشند. آنها برای آزمودنی i ام $(i = 1, \ldots, m)$، تابع چگالی توام پاسخ‌ها را به صورت زیر درنظر می‌گیرند:

$$f(y_i ; \Psi_i , \Omega_i) = \exp\{\Psi'_i y_i + \Omega'_i W_i - A(\Psi_i, \Omega_i)\} \qquad (۲)$$

که درآن:

$$W_i = (y_{i1}y_{i2}, \ldots, y_{in-1}y_{in}, \ldots, y_{i1}y_{i2}\cdots y_{in})$$

برداری $1 \times (2^n - n - 1)$ از حاصلضرب‌های دوتایی و بالاتر y_i و $\Psi_i = (\psi_{i1}, \ldots, \psi_{in})'$ و $\Omega_i = (\omega_{i12}, \ldots, \omega_{i(n-1)n}, \ldots, \omega_{i12\ldots n})'$ بردارهایی از پارامترهای متعارف می‌باشند. $\exp\{-A(\Psi_i, \Omega_i)\}$ مقدار ثابتی است که به y_i وابسته نیست.

۴- رهیافت‌های حاشیه‌ای دیگر براساس درستنمایی

دراین بخش مدل‌های حاشیه‌ای براساس درستنمایی ارائه می‌شوند. درتمام این مدل‌ها، توزیع توام پاسخ‌ها بوسیله درنظر گرفتن فرض‌هایی به طور کامل مشخص می‌شوند. یکی‌از این فرض‌ها درباره ارتباط پاسخ‌ها است. دراین بخش رهیافت‌های حاشیه‌ای که معیارهای همبستگی آنها مقید به احتمال‌های حاشیه‌ای نیستند _ لگ نسبت بخت شرطی، نسبت وابستگی و لگ نسبت بخت حاشیه‌ای _ بیان می‌شوند.

درتمام این مدل‌ها ارتباط پاسخ‌ها با متغیرهای کمکی بصورت $\log it(\mu_{ij}) = x'_{ij}\beta$ و واریانس پاسخ‌ها بصورت $\mathrm{var}(Y_{ij}) = \mu_{ij}(1-\mu_{ij})$ بیان می‌شود. تنها تفاوت آنها درمعیار همبستگی پاسخ‌ها است. توجه داشته باشید که لوجیت $(\log it)$ به‌معنای لگاریتم بخت است یعنی:

$$\log it(\mu_{ij}) = \log\left(\frac{P(Y_{ij}=1)}{1-P(Y_{ij}=1)}\right)$$

۴-۱) مدل حاشیه‌ای با پارامترهای همبستگی به صورت توابعی از لگ نسبت بخت شرطی

دراین مدل‌بندی که توسط فیتزموریس و لرد پیشنهاد شده است (Fitzmaurice & Laird, 1993). در معادله ۲ همبستگی پاسخ‌ها برای هر آزمودنی براساس لگ نسبت بخت شرطی و توابعی از آنها قابل بیان است.

درمعادله۲ پارامترهای Ψ_i عباراتی از احتمال‌های شرطی به صورت:

$$\psi_{ij} = \log it\{\Pr(y_{ij} - 1)\}$$

و پارامترهای Ω_i (تعریف‌شده درمعادله۲)عباراتی از لگ نسبت بخت شرطی‌است یعنی:

$$\exp(\omega_{irs}) = \frac{\Pr(Y_{ir} = 1, Y_{is} = 1 | Y_{it} = ۰, t \neq r,s)}{\Pr(Y_{ir} = 1, Y_{is} = ۰ | Y_{it} = ۰, t \neq r,s)} \times \frac{\Pr(Y_{ir} = ۰, y_{is} = ۰ | Y_{it} = ۰, t \neq r,s)}{\Pr(Y_{ir} = ۰, y_{is} = 1 | Y_{it} = ۰, t \neq r,s)}$$

و باید توجه کرد که μ_i تابعی‌از هر دو بردار Ψ_i و Ω_i است.

دراین مدل پارامترهای همبستگی مقید به میانگین حاشیه‌ای نیستند و فضای پارامترها برای $(\omega_{i12} \ \ \omega_{i12\ n})'$ $\Omega_i =$ ، $\Re^{2^n - n - 1}$ (فضای اقلیدسی۱-n- " ۲ بعدی) می‌باشد. ولی این نوع پارامترها به n (تعداد پاسخ‌های تکرار شده برای هر آزمودنی) وابسته می‌باشند. یعنی می‌توان گفت این نوع مدل‌بندی برای داده‌هایی با تعداد پاسخ‌های تکرار شدهٔ متفاوت برای برخی‌از آزمودنی‌ها قابل کاربرد نیستند.

۲-۴) مدل حاشیه‌ای با پارامترهای همبستگی به صورت توابعی از نسبت وابستگی

این مدل حاشیه‌ای توسط اخولم و دیگران، پیشنهاد شده است (Ekholm *et.al.*, 1995). در این روش استفاده از پارامتر میانگین (مولفه‌هایی که احتمال موفقیت توام تمام مراتب پاسخ‌ها را دربردارد) برای تحلیل رگرسیون پاسخ دودویی چند متغیره پیشنهاد شده‌است که در آن همبستگی را با استفاده از نسبتهای وابستگی (که عبارت‌هایی از پارامتر میانگین می‌باشد) بیان می‌کنند.

اگر $S(Y_i) = (Y_i, W_i)'$ را درنظر بگیریم که Y_i و W_i همان بردارهای تعریف شده درمعادله ۲ هستند آنگاه به بردار مقادیر مورد انتظار $S(Y_i)$، پارامتر میانگین_ ϕ_i _می‌گویند (Barndorff-Neilson & Cox, 1994).

یعنی:

$$\phi_i = E\{S(Y_i)\} = (\phi_{i1}, \ldots, \phi_{in}, \phi_{i12}, \ldots, \phi_{i(n-1)n}, \ldots, \phi_{i12\ldots n})'$$

مولفه‌های ϕ_i احتمال‌های موفقیت توام حاشیه‌ای را برای اندیس‌های آن بیان می‌کند یعنی:

$$\phi_{i1} = \Pr(Y_{i1} = 1), \ldots, \phi_{i12} - \Pr(Y_{i1} = 1, Y_{i2} = 1), \ldots,$$

$$\phi_{i12\ldots n} = \Pr(Y_{i1} = 1, \ldots, Y_{in} = 1)$$

برای مثال پاسخ دودویی دومتغیره برای آزمودنی i ام، (Y_{i1},Y_{i2})، با پارامتر میانگین $\phi_i=(\phi_{i1},\phi_{i2},\phi_{i12})'$ را در نظربگیرید. نسبت وابستگی، ω_{i12}، و لگ نسبت وابستگی، λ_{i12}، به صورت زیر تعریف می‌شود:

$$\phi_{i12\ldots n}=\Pr(Y_{i1}=1,\ldots,Y_{in}=1)$$

$$\omega_{i12}=\frac{\phi_{i12}}{\phi_{i1}\phi_{i2}},\ \lambda_{i12}=\log(\omega_{i12}) \qquad (۳)$$

که درآن ۱۰ × (۱ - ω_{i12}) میزان درصد بزرگی احتمال موفقیت توام Y_{i1} و Y_{i2} درمقایسه با فرض استقلال را نشان می‌دهد.

نسبت‌های وابستگی مراتب بالاتر نیز به صورت زیر تعریف می‌شود:

$$\omega_{i123}=\frac{\phi_{i123}}{\phi_{i1}\phi_{i2}\phi_{i3}},\ldots,\omega_{i123\ldots n}=\frac{\phi_{i123\ldots n}}{\phi_{i}\phi_{i2}\cdots\phi_{in}}.$$

نسبت‌های وابستگی بوسیله مقایسه احتمال‌های موفقیت حاشیه‌ای با حالت استقلال تفسیرمی‌شوند، اگرنسبت وابستگی با دو اندیس و بیشتر، برابر با یک باشد- برای نمونه اگر $\omega_{i123}=1$ آنگاه $\phi_{i123}=\phi_{i1}\phi_{i2}\phi_{i3}$ که استقلال سه‌تایی $Y_{i3}Y_{i2}Y_{i1}$ را نتیجه می‌دهد. باید توجه‌داشت‌که اگر موفقیت و شکست برای مولفه‌هایی دو دویی معاوضه شوند (مثلا برای Y_{i2}) و ($1-Y_{i2}$ و Y_{i1}) مطالعه شود آنگاه ضریب همبستگی و لگ نسبت بخت درعلامت تغییر می‌کنند اما در مقدار قدر مطلق تغییر نمی‌کنند اما ω_{i123} به $\frac{1-\phi_{i12}\omega_{i12}}{1-\phi_{i12}}$ تبدیل می‌شود. علاوه براین ضریب همبستگی ونسبت‌های بخت برای ($1-Y_{i2}$ و $1-Y_{i1}$) برابر با (Y_{i2},Y_{i1}) است درحالی که نسبت وابستگی برای ($1-Y_{i2}$ و $1-Y_{i1}$) تنها تابعی صعودی از نسبت وابستگی برای (Y_{i2},Y_{i1}) است.

۳ـ۴) مدل حاشیه‌ای با لگ نسبت بخت حاشیه‌ای به عنوان پارامتر همبستگی:

این مدل‌بندی توسط لیانگ و دیگران و لیپسیتز و دیگران ارائه شده است (Liang *et al., 1992;* Lipsitz *et a., 1991)*. در این مدل‌بندی، درمعادله۳ پارامترهای Ψ_i عباراتی از احتمال‌های حاشیه‌ای می‌باشند یعنی:

$$\psi_{it} = \log it\{\Pr(Y_{it}=1)\}$$

و پارامترهای دو طرفه Ω_i لگ نسبت بخت حاشیه‌ای می‌باشند یعنی:

$$\exp(\omega_{irs}) = OR(Y_{ir}, Y_{is}) = \frac{\Pr(Y_{ir}=1, Y_{is}=1)\Pr(Y_{ir}=0, Y_{is}=0)}{\Pr(Y_{ir}=0, Y_{is}=1)\Pr(Y_{ir}=1, Y_{is}=0)}$$

و پارامترهای سه طرفه و بالاتر به صورت توابعی از لگ نسبت بخت حاشیه‌ای است. به عنوان مثال:

$$\exp(\omega_{irst}) = \frac{OR(Y_{ir}, Y_{is} | Y_{it}=1)}{OR(Y_{ir}, Y_{is} | Y_{it}=0)}$$

یا بطور معادل:

$$\omega_{irst} = \log[OR(Y_{ir}, Y_{is} | Y_{it}=1)] - \log[OR(Y_{ir}, Y_{is} | Y_{it}=0)]$$

است.

با این نوع مدل‌بندی‌ها، درجه‌های مختلف وابستگی درمیان Y_{it}ها قابل بیان است. به عنوان مثال مدل استقلال نتیجه می‌دهد که تمام پارامترهای همبستگی دو طرفه و بالاتر صفر است و یا اگر تمام عناصر Ω_i مخالف صفر باشند مدل اشباع شده می‌باشد. بین این دو حالت مدل‌های مختلفی از وابستگی را می‌توانیم درنظر بگیریم. همانطورکه ملاحظه می‌شود اینگونه مدل‌ها محاسبات زیادی برای برآورد پارامترها دارند (بجزبرای $n=2$). درنتیجه برای رهایی‌ازاین مشکل محققین اکثراً ارتباطهای مراتب بالاتر از دو را صفر درنظر می‌گیرند(Zhao & Prentice, 1990).

۵- مدل‌های حاشیه‌ای با استفاده از مفهوم متغیر پنهان

از مدل پروبیت چند متغیره با استفاده از مفهوم متغیر پنهان (Latent variable) برای تحلیل‌های طولی دو دویی استفاده می‌شود (Long, 1997) چون Y_{it}^* متغیری پیوسته است دسته‌ای از مدل‌های خطی را می‌توان برای تحلیل آنها درنظرگرفت، که به صورت زیر تعریف می‌شوند.

$$Y_{it}^* = x'_{it}\beta + \varepsilon_{it}$$

که درآن فرض‌هایی درباره ε_{it}ها درنظر گرفته می‌شود (Long, 1997). به عنوان مثال:

$\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in})'$ برای آزمودنی iام، دارای توزیع نرمال n متغیره با میانگین صفر و ماتریس کوواریانس:

$$\Sigma=\begin{pmatrix}1 & \rho_{12} & \dots & \rho_{1n}\\ \rho_{21} & 1 & \dots & \rho_{2n}\\ & & \ddots & \\ \rho_{n1} & \dots & \dots & 1\end{pmatrix}=\rho$$

درنظر گرفته می‌شود. این فرض مدل پروبیت چند متغیره را نتیجه می‌دهد.

با توجه به این‌که y_{it}های مشاهده شده بر اساس قاعده زیر مشخص می‌شوند.

$$y_{it}=\begin{cases}1 & y_{it}^*>\tau\\ \circ & y_{it}^*\le\tau\end{cases}$$

که $\tau=\circ$ اختیار می‌شود (Long, 1997)، احتمال‌های توام و درنتیجه تابع درستنمایی را با استفاده از فرض‌هایی روی ε، می‌توان محاسبه نمود.

همانطور که ملاحظه می‌شود با افزایش n تعداد پارامترهای ماتریس همبستگی افزایش چشمگیری خواهد داشت، $\frac{n(n-1)}{2}$ پارامتر، که برآورد تمام این پارامترها نیاز به محاسبات و وقت زیاد دارد. به همین دلیل برای ماتریس همبستگی، حالت‌های خاصی درنظر گرفته می‌شود. به عنوان مثال:

$$Corr(\varepsilon_{it},\varepsilon_{it'})=\rho,\ \forall t\neq t' \qquad (1)$$

یا

$$Corr(\varepsilon_{it},\varepsilon_{it'})=\rho^{|t-t'|} \qquad (۲)$$

که درحالت۲ همبستگی بین دو پاسخ آزمودنی یکسان با افزایش فاصله زمانی بین دو پاسخ کاهش می‌یابد.

۶- رهیافت معادلات برآورد تعمیم یافته

این رهیافت توسط لیانگ و زیگرپیشنهاد شده است (Liang and Zeger, 1981). رهیافت معادلات برآورد تعمیم یافته، رهیافت واحدی را برای تحلیل برآمدهای گوناگون پیوسته و گسسته پیشنهاد می‌کند. دراین میان دسته‌ای از معادلات برآوردگر تعمیم‌یافته (GEE) برای پارامترهای رگرسیون پیشنهاد شده‌است. دراین رهیافت تنها فرض‌هایی‌که اختیار می‌شود فرض درباره امید حاشیه‌ای مرتبه اول و دوم پاسخ‌ها است و درباره توزیع کامل آنها فرضی اختیار نمی‌شود.

این رهیافت برای پارامترهای رگرسیون برآوردهای سازگار می‌دهد. دراصل (GEE) حالت چند متغیره شبه درستنمایی (Quasi likelihood) است که توسط ودربرن (Wedderburn، 1974) بیان شده‌است.

دراین رهیافت تابعی‌از امید حاشیه‌ای متغیر وابسته به صورت تابعی خطی‌از متغیرهای کمکی مشخص می‌شود. همچنین فرض می‌شود که واریانس تابعی معلوم از میانگین حاشیه‌ای است. به علاوه ماتریس همبستگی عملی برای مشاهدات هر آزمودنی، معین می‌شود. یعنی:

$$var\ (Y_{ij})=v(\mu_{ij})/\phi \quad و \quad h\ (\mu_{ij})=x'_{ij}\beta$$

و همبستگی عملی مشاهدات برای آزمودنیi ام $R_i\ (\alpha)$ فرض می‌شود که به بردار $s\times 1$ از پارامترهای اضافی، α، وابسته است. α، همبستگی پاسخ‌های آزمودنیi ام را به طور کامل بیان می‌کند. اکثراً از $R\ (\alpha)$ به عنوان ماتریس همبستگی عملی استفاده می‌شود، زیرا انتظار این که ماتریس همبستگی واقعی به درستی مشخص شود نیست. درنتیجه ماتریس کوواریانس عملی Y_i به صورت:

$$V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2} / \phi$$

است که A_i ماتریس قطری n_i، n_i با عناصر$v\ (\mu_{ij})$ درjامین عنصر قطری آن می‌باشد.
معادلات برآورد به صورت زیر می‌باشد.

$$\sum_{i=1}^{m} D_i' V_i^{-1} S_i = 0 \qquad (۴)$$

که در آن $S_i=Y_i-\mu_i$ و $D_i=\dfrac{\partial\mu_i}{\partial\beta}$ است. هنگامی که $n_i=1$ ، $\forall_i=1, \ldots, m$ معادله (۴) به معادله شبه درستنمایی تحویل می‌شود (Wedderburn, 1974).

معادله۴ می‌تواند به صورت تنها تابعی‌از β بیان شود به‌این صورت که برآوردگرهای سازگار α و ϕ درمعادله ۴ قرارگیرد (لیانگ و زیگر، ۱۹۸۶) یعنی به صورت زیر خواهد بود.

$$\sum_{i=1}^{m} U_i\left[\beta,\hat{\alpha}\left\{\beta,\hat{\phi}(\beta)\right\}\right]=0 \qquad (۵)$$

$\hat{\beta}_G$ که نشان دهنده برآورد β تحت رهیافت (GEE) است، جواب معادلهٔ بالا خواهد بود. براساس قضیه۲ در لیانگ و زیگر (۱۹۸۶)، $m^{1/2}\left(\hat{\beta}_G-\beta\right)$ بطور مجانبی گاوسی با میانگین صفر و ماتریس کوواریانس V_G به صورت زیر است:

$$
\begin{aligned}
V_G &= \lim_{m\to\infty} m\left(\sum_{i=1}^{m} D_i' V_i^{-1} D_i\right)^{-1} \\
&\times \left\{\sum_{i=1}^{m} D_i' V_i^{-1} \operatorname{cov}(Y_i) V_i^{-1} D_i\right\}\left(\sum_{i=1}^{m} D_i' V_i^{-1} D_i\right)^{-1} \\
&= \lim_{m\to\infty} m V_1^{-1} V V_1^{-1}
\end{aligned}
$$

برآورد واریانس $\hat{\beta}_G$ _ $\hat{V}_G$ _ می‌تواند برای یافتن واریانس برآوردگرها استفاده شود (Zeger & Liang, 1992).

۷- مقایسه کلی مدل‌های حاشیه‌ای

در این بخش به بررسی و مقایسه مدل‌های حاشیه‌ای ارائه شده دربخش‌های قبل می‌پردازیم. ابتدا مدل استقلال، که فرض برعدم همبستگی بین پاسخ‌های متعلق یک آزمودنی خاص دارد، را در نظر گیرید. براساس قضیه ۱ لیانگ و زیگر (۱۹۸۶)، $\hat{\beta}_I$ (برآورد پارامتر تحت مدل استقلال) برآوردی سازگار برای β مهیا می‌کند. اما این روش استنباط‌هایی نادرست درباره پارامترها می‌دهد زیرا فیتزموریس و دیگران (۱۹۹۳) بیان کردند که اکثرا خطای استاندارد برآورد پارامترهای مربوط به متغیرهای زمان مانا میل به کم برآورد شدن و برای پارامترهای مربوط به متغیرهای زمان متغیر میل به بیش برآورد شدن دارند. و بنابراین برآوردهای واریانس‌ها سازگار نیستند.

رهیافت‌هایی که دربخش چهارم بیان شدند هرکدام دارای محاسن و معایبی می‌باشند. پارامترهای همبستگی به صورت توابعی از لگ نسبت بخت شرطی بر خلاف همبستگی‌های بهادر و لگ نسبت بخت حاشیه‌ای مقید به میانگین حاشیه‌ای نیستند. ولی این نوع پارامترها به n (تعداد پاسخ‌های تکرار شده برای هر آزمودنی) وابسته می‌باشند.

وقتی پارامترهای همبستگی به صورت توابعی از لگ نسبت بخت حاشیه‌ای بیان شود، پارامترهای همبستگی به n وابسته نیست. و به‌همین دلیل برای تعداد پاسخ‌های تکرار شده متفاوت برای آزمودنی‌های مختلف، این مدل کاربرد دارد. دیگر مزیت مدل لگ نسبت بخت حاشیه‌ای نسبت به مدل با لگ نسبت بخت شرطی آن است که قابل دوباره تولید شدن می‌باشد ولی مدل با لگ نسبت بخت شرطی قابلیت دوباره تولید شدن را ندارد (مدلی قابل دوباره تولید شدن است که اگر $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{in})'$ دریک مدل حاشیه‌ای صدق کند، آنگاه زیر مجموعه‌ای از Y_i مانند $Y_i' = (Y_{i1}, \dots, Y_{in'})'$، $(n' < n)$ نیز درمدل با پارامترهای متناظر آن صدق کند).

هنگامی‌که پارامترهای همبستگی بین پاسخ‌ها براساس نسبت وابستگی بیان می‌شود، پارامترهای همبستگی همانند لگ نسبت بخت حاشیه‌ای وابسته به n نیست. باید توجه داشت که پیچیدگی قید پارامترهای همبستگی نسبت‌های وابستگی ولگ نسبت بخت حاشیه‌ای به میانگین‌ها کمتر از پیچیدگی قید همبستگی‌های بهادر است.

در هرسه مدل بخش۴، ضرایب رگرسیونی و پارامترهای همبستگی بر هم عمود می‌باشند. این بدین معنی است که کئواریانس برآوردگرهای ضرایب رگرسیونی و پارامترهای همبستگی برابر صفر است. (بارندورف نیلسون و کاکس، ۱۹۹۴).

برآوردهای ماکسیمم درستنمایی پارامترهای مدل‌های ذکر شده در بخش چهارم براساس روشهای تکراری عددی بدست می‌آیند. برای تمام این مدل‌ها بدون توجه به ساختار همبستگی، برآورد پارامترهای رگرسیونی سازگار می‌باشند ولی تنها اگر ساختار همبستگی برای این مدل‌ها بدرستی مشخص شده باشند آنگاه برآورد $\mathrm{cov}(\hat{\beta})$ نیز برآوردی سازگار می‌باشد.

رهیافت GEE که رهیافتی غیردرستنمایی‌است برخلاف مدل پروبیت چند متغیره که به فرض توزیع چند متغیره نرمال برای متغیرهای پنهان نیاز دارد، تنها به فرضهایی درباره دو گشتاور اول پاسخ‌ها نیاز دارد. البته باید توجه کرد که در توزیع گاوسی، فرض درباره دو گشتاور اول، توزیع را به طور کامل مشخص می‌کند درحالی که برای داده‌های دودویی که مورد بحث اصلی است، این مطلب برقرار نمی‌باشد (توجه کنید که متغیرهای پنهان ما دارای توزیع گوسی‌اند نه متغیرهای مشاهده شده). درGEE تابع پیوندهای مختلفی همچون پروبیت، لوجیت و ... می‌تواند مورد استفاده قرار گیرد، ولی در پروبیت چند متغیره تنها پیوند پروبیت و در مدل برای لگ نسبت بخت حاشیه‌ای تنها پیوند لوجیت مورد استفاده است. GEE همانند مدل با نسبت وابستگی و نسبت بخت حاشیه‌ای، برای داده‌های با اندازه تکرارهای متفاوت برای آزمودنی‌های مختلف قابل اجرا است.

یکی از محدودیتهای GEE نداشتن روش‌هایی برای آزمون نیکویی برازش، مقایسه مدل و استنباط بر روی پارامترها بر اساس درستنمایی است زیرا در این روش، توزیع توام به طور کامل مشخص نمی‌شود. با توجه به این که برآوردگرها دارای توزیع نرمال مجانبی هستند، در این روش برای حل این مشکل از آماره‌های والد (Wald Statistics) استفاده می‌شود. یکی دیگر از محدودیتهایGEE آن است که اگر مدل برای میانگین‌های حاشیه‌ای به درستی مشخص نشده باشد، برآوردهای ناسازگار برای β به دست می‌آید.

۸- بررسی مانده‌ها

بررسی مانده‌ها برای نیکویی برازش مدل و یافتن دورافتاده‌ها (مشاهداتی که روی نیکویی برآزش مدل تاثیر می‌گذارند) درحالت مطالعات طولی، همانند مطالعات مقطعی ضروری است. تفاوتی که در این حالت برای مانده‌ها نسبت به مطالعات مقطعی وجود دارد این است که برای هر آزمودنی چند پاسخ داریم. برای محاسبه مانده‌ها اگر $\Sigma_i(\beta) = diag\{\mathrm{var}(Y_{it})\}$ درنظر گرفته شود آنگاه همبستگی بین پاسخ‌های متعلق به یک آزمودنی خاص درنظر گرفته نمی‌شود. اما اگر از ماتریس کوواریانس Y_i استفاده شود، این همبستگی برای محاسبه مانده‌ها نیز مورد استفاده قرار می‌گیرد. در این حالت مانده‌های پیرسون برای مشاهدات Y_i را به صورت زیر معرفی می‌کنیم. مانده‌ها به صورت:

$$r_i^{p} = \sum\nolimits_i^{-1/2}(\hat{\beta})(y_i - \hat{\mu}_i)$$

داده می‌شود که درآن $\sum_i(\hat{\beta})$ برآورد ماتریس کوواریانس مشاهدات است. همچنین آماره نیکویی برازش پیرسون به صورت:

$$\chi^2 = \sum_{i=1}^{m} \chi_p^2(y_i, \hat{\mu}_i)$$

است که درآن:

$$\chi_p^2(y_i, \hat{\mu}_i) = (y_i - \hat{\mu}_i)' \Sigma_i^{-1}(\hat{\beta})(y_i - \hat{\mu}_i)$$

$\chi_p^2(y_i, \hat{\mu}_i)$ را می‌توان به صورت: $r_i^{p'} r_i^{p}$ که درآن $r_i^{p'}$ ترانهاده بردار r_i^{p} تعریف شده دربالا است، نوشت. بزرگ بودن χ_p^2 درمقایسه با درجه‌آزادی آن (تعداد پارامترهای برآورد شده- m) گواه بر بدی برآزش مدل دارد. دراین صورت یا مدل مورد استفاده مناسب نیست یا، درصورت وجود، مؤلفه‌های دیگری بایستی درمدل به‌عنوان متغیرهای تبیینی درنظر گرفته شود.

به‌عنوان مثال با مدل دو متغیره پروبیت با دو پاسخ برای هر آزمودنی به جای ساختار کوواریانس زیر که فرض استقلال را درنظر می‌گیرد:

$$\Sigma_i(\beta) = \begin{bmatrix} \pi_{i1}(1-\pi_i) & 0 \\ 0 & \pi_{i2}(1-\pi_{i2}) \end{bmatrix}$$

ماتریس کوواریانس Y_i به صورت زیر را در نظر می‌گیریم:

$$\sum\nolimits_i(\beta) = \begin{bmatrix} \pi_{i1}(1-\pi_{i1}) & \pi_{i12} - \pi_{i1}\pi_{i2} \\ \pi_{i12} - \pi_{i1}\pi_{i2} & \pi_{i2}(1-\pi_{i2}) \end{bmatrix}$$

که درآن:

$$cov(Y_{i1}, Y_{i2}) = \pi_{i12} - \pi_{i1}\pi_{i2}$$

با:

$$\pi_{i12} = \Pr(Y_{i1} = 1 . Y_{i2} = 1)$$
$$= 1 - \Phi(-x'_{i1}\beta_1) - \Phi(-x'_{i2}\beta_2) + \Phi_2(-x'_{i1}\beta_1, -x'_{i2}\beta_2, \rho)$$

و

$$\pi_{ij} = \Pr(Y_{ij} = 1) = \Phi(x'_{ij}\beta) \qquad j = ۱ و ۲$$

$\Phi_2(\cdot, \ldots, \rho)$ تابع توزیع نرمال دو متغیره با همبستگی ρ بین دو متغیر است و$\Phi(\cdot)$ تابع توزیع نرمال استاندارد است. اگر برآورد β را در $\sum_i^{-1/2}(\beta)$ قرار دهیم برآورد این ماتریس بدست می‌آید. همان طورکه ملاحظه می‌شود برآورد پارامتر ρ در محاسبه برآورد ماتریس کوواریانس و بنابراین مانده‌ها نقش دارد.

۹- مثال عملی: (برازش مدل‌های حاشیه‌ای بر روی داده‌های آسم)

این داده‌ها (رتنیتزکی و وایپیچ، ۱۹۹۴) در۶ شهرستان منطقه *هاروارد، اهایو* جمع‌آوری شده‌اند، به این ترتیب که ۷۰۶ پسر و ۷۱۳ دختر سفید پوست را در ۹ سالگی و دوباره در ۱۳ سالگی از نظر ابتلا یا عدم ابتلا به آسم مورد مطالعه قرار داده‌اند. می‌خواهیم تاثیر زمان و جنس را بر احتمال داشتن آسم مورد آزمون قرار دهیم. جدول۱ داده‌های کامل و بدون مقادیر گمشده برای متغیر پاسخ در این داده‌ها را نشان می‌دهد. در بررسی‌های جدا برای رد آزمون کاملا تصادفی بودن مقادیر گمشده گواهی بدست نیامد (Little and Rubin, 1987). همانطورکه جدول۱ نشان می‌دهد ۵۵۷ نفر پسر و ۵۹۰ نفر دختر به متغیر مورد علاقه پاسخ داده‌اند. برای بررسی تاثیر سن و زمان بر احتمال داشتن آسم، مدل‌های حاشیه‌ای که قبلا ذکر شد را برازش می‌دهیم.

مدل برای نسبت وابستگی، لگ نسبت بخت حاشیه‌ای و مدل استقلال به صورت زیر در نظر گرفته شده است:

$$\log it(\pi_{ij}) = a_j + b_j sex_i \quad j = ۱ و ۲ \quad . \quad i = 1, \ldots, ۱۱۴۷$$

که در آن π_{ij} احتمال داشتن آسم فرد i ام در زمان j ام است، sex_i جنس آزمودنی i ام (۱ : پسر، ۰ : دختر) را نشان می‌دهد و $\text{var}(Y_{ij}) = \pi_{ij}(1 - \pi_{ij})$ است. پارامتر همبستگی برای نسبت وابستگی با لگ نسبت وابستگی و با لگ نسبت بخت حاشیه‌ای برای

مدل نسبت بخت حاشیه‌ای بیان می‌شود. مدل حاشیه‌ای پروبیت دو متغیره با مفهوم متغیر پنهان نیز به این داده‌ها برازش داده‌ایم. مدل در نظر گرفته شده به صورت:

$$Y_{ij}^* = a_j + b_j sex_i + \varepsilon_{ij} \qquad j=۱و۲$$

است. که Y_{i1}^* و Y_{i2}^* متغیرهای پنهان هستند. همچنین همانطور که دربخش ۵ ذکر شد $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2})'$ برداری با توزیع نرمال دو متغیره است.

جدول ۱- داده‌های مربوط به آسم

	۱۳سالگی				
جمع کل	اسم ندارد	آسم دارد			
۲۸	۶	۲۲	آسم دارد	**پسرها**	**۹ سالگی**
۵۲۹	۵۱۴	۱۵	آسم ندارد		
۵۵۷	۵۲۰	۳۷	جمع کل		
۱۶	۳	۱۳	آسم دارد	**دخترها**	
۵۷۴	۵۶۱	۱۳	آسم ندارد		
۵۹۰	۵۶۴	۲۶	جمع کل		

نتایج برازش مدل‌ها در جدول۲ آمده است. با استفاده از نتایج بدست آمده در جدول۲ مشاهده می‌شود که پارامترهای همبستگی وجود همبستگی بالای مثبت(ρ - مقدار = ۰/۰۰۰) را بین دو پاسخ برای هر آزمودنی نشان می‌دهد.

همان‌گونه‌که ملاحظه می‌شود تمام این مدل‌ها ضریب ثابت را معنی‌دار نشان می‌دهند یعنی عواملی مانند زمان بوده‌اند که درنظر نگرفته‌ایم. به‌این دلیل دوباره مدلGEE را بر روی داده‌ها برازش داده‌ایم. مدل GEE به صورت زیر تعیین می‌گردد:

$$\operatorname{logit}(\pi_{ij}) = a_1 + b_1 sex_j + cI(age13)_{ij} \qquad j=۱,۲$$

که درآن I (age 13) به صورت زیر تعریف می‌شود:

$$I(age13) = \begin{cases} \text{اگر فرد سیزده ساله باشد،} & ۱ \\ \text{اگر فرد سیزده ساله نباشد،} & ۰ \end{cases}$$

در GEE ماتریس همبستگی به صورت غیر ساختاری درنظر گرفته شده است. نتایج برازش مدل‌ها درجدول۲ آمده است.

هرچند مقایسه مدلها دریک مثال امکان پذیر نیست و نیاز به بررسی‌های نظری الزامی است، خلاصه‌ای از عملکرد مدل‌های مختلف در برآورد پارامترها می‌آوریم. درهمه مدلها گواه کافی بر معنی‌داری پارامتر منتسب به زمان (c یا تفاضل a_2 , a_1) وجود دارد. به این معنی که هر چه زمان اقامت دراین ۶ شهرستان بیشتر می‌شود احتمال ابتلا به آسم افزایش می‌یابد. در مدل حاشیه‌ای پروبیت دو متغیره جنس در زمان اول معنی‌دار است ولی در زمان دوم معنی‌دار نیست. به این معنی‌که پسران درسنین پایین برای آسم داشتن محتمل‌تر از دختران هستند ولی با گذشت زمان پسران بهبود یافته بیشتر از دخترانی هستند که به آسم مبتلا می‌شوند. مدل پروبیت با فرض استقلال نیز به نتیجه یکسانی می‌رسد، اما واریانس‌های برآوردگرهای پارامترها کمی بیش یا کم برآورد می‌شوند. مدل استقلال با پیوند لوجیت نیز در مورد برآورد پارامترها به نتیجه‌ای شبیه به مدل پروبیت دو متغیره می‌رسد، ولی این مدل در مقایسه با مدل نسبت وابستگی با پیوند لوجیت یا نسبت بخت با پیوند لوجیت واریانس برآوردگرها را بسیار متزلزل برآورد می‌کند.

جدول۲- برآورد پارامترها تحت مدلهای در نظر گرفته شده.

مدل استقلال با پیوند لوجیت		نسبت وابستگی با پیوند لوجیت		نسبت بخت با پیوند لوجیت		پارامتر
خطای استاندارد	برآورد	خطای استاندارد	برآورد	خطای استاندارد	برآورد	
۰/۲۵۳	-۳/۵۸۰	۰/۲۳۲	-۳/۳۹۰	۰/۴۱۸	-۵/۰۶۰	a_1
۰/۲۰۰	-۳/۰۷۷	۰/۱۷۱	-۲/۸۸۸	۰/۲۶۰	-۳/۷۲۳	a_2
۰/۳۱۹	۰/۶۴۱	۰/۲۶۲	۰/۲۷۸	۰/۴۱۳	۰/۵۲۴	b_1
۰/۲۶۳	۰/۴۳۴	۰/۱۸۶	۰/۰۷۳	۰/۳۴۲	۰/۱۵۳	b_2
----	----	۱/۶۴۵	۱۴/۲۴۱	۰/۴۲۱	۴/۹۸۶	پارامتر همبستگی
۴۲۷/۱۸۱		۳۳۸/۷۴۶		۳۳۷/۲۱۱		منهای لگاریتم درستمایی
مدل استقلال با پیوند پروبیت		**مدل پروبیت**		**مدل GEE با پیوند لوجیت**		**پارامتر**
خطای استاندارد	برآورد	خطای استاندارد	برآورد	خطای استاندارد	برآورد	
۰/۱۰۶	-۱/۹۲۷	۰/۱۰۵	-۱/۹۲۵	۰/۲۱۱	-۳/۴۸۶	a_1
۰/۰۹۱	-۱/۷۰۵	۰/۰۹۰	-۱/۷۰۵	----	----	a_2
۰/۱۳۹	۰/۲۸۳	۰/۱۳۵	۰/۲۸۳	۰/۲۵۵	۰/۴۸۹	b_1
۰/۱۲۲	۰/۲۰۲	۰/۱۲۱	۰/۲۰۲	----	----	b_2
----	----	----	----	۰/۱۲۱	۰/۳۷۷	c
----	----	۰/۲۴۵	۰/۹۲۶	----	۰/۶۴۴	پارامتر همبستگی
۴۲۷/۱۸۱		۳۳۷/۱۰۶		------		منهای لگاریتم درستمایی

در مدلهای نسبت وابستگی با پیوند لوجیت یا نسبت بخت با پیوند لوجیت اثر جنس در هیچیک از زمانها معنی‌دار نیست، به این معنی‌که پسرها و دخترها درهر سنی به یک‌اندازه برای داشتن آسم محتملند. درمدل GEE، که درآن بر خلاف مدلهای دیگر هیچ فرضی درمورد توزیع پاسخها اعمال نمی‌شود، گواهی قوی بر معنی‌داری اثر زمان ولی گواهی ضعیف بر معنی‌داری جنس وجود دارد.

- برای نیکویی برازش مدل بر اساس مانده‌ها که در بخش ۸ بیان شده است، می‌توانیم عمل کنیم. مثلا برای مدل پروبیت دو متغیره برازش شده بر روی این داده‌ها، آماره کی دو برابر با ۲۵۴۳/۰۸۹ است که با این مقدار آماره فرض مناسب بودن مدل رد می‌شود. اگر در مدل پروبیت دو متغیره همبستگی را در نظر نگیریم مقدار آماره برابر با ۲۵۵۸/۰۸۴ است و خوبی برآزش این مدل نیز رد می‌شود. فکر می‌کنیم که عدم برازش خوب به دلیل وجود عاملهای متعدد موثر دیگری همچون میزان آلودگی هوا و عوامل محیطی و ژنتیکی دیگر است، که دراین بررسی درنظر گرفته نشده است. بهر حال کاهش مقدار آماره کی دو درمدل عدم استقلال نسبت به مدل استقلال به خاطر درنظر گرفتن همبستگی بین پاسخ‌ها است.

Reference

Agresti, A. (2002) *Categorical Data Analysis*. New York: John Wiley.

Bahadur, R.R. (1961) *A representation of the joint distribution of responses to n dichotomous item. In studies on item analysis and prediction.* Stanford: Standorf University Press.

Barndorff-Nielson, O.E., and Cox, D.R. (1994) *Inference and Asymptotics*. London: Chapman & Hall.

Diggel, P.J., Liang, K., Zeger, S. (1994) *Analysis of Longitudinal Data*. Oxford Science Publication.

Ekholm, A., Smith, P.W.F., McDonald, J.W. (1995) Marginal regression analysis of a multivariate binary response. *Biometrika*, **82**,847-854.

Fitzmaurice, G.M., Laird, N.M. (1993) *A Likelihood-Based method for analysing longitudinal binary responses*. Biometrika, **80**, 141-151.

Fitzmaurice, G.M., Laird, N.M. Rotnitzky, A.G. (1993) *Regression Models for Discrete Longitudinal responses*. Statistical Science, **8**, 284-309.

Liang, K.Y., Zeger, S.L. (1986) *Longitudinal data analysis using generalized linear models.* Biometrika, **73**, 13-22.

Liang, K.Y., Zeger, S.L., Qaqish, B. (1992) *Multivariate regression analysis for categorical data (with discussion).* Journal of the Royal Statistical Society, B, **54**, 3-40.

Lipsitze, S., Laird, N., Harrington, D. (1991) *Generalized estimating equations for corrolated binary data: Using odds ratios as a measure of association.* Biometrika, **78**, *153-160.*

Little, R. J. and Rubin, D. (1987) *Statistical analysis with missing data.* New York: Wiely. .

Long, S.J. (1997) *Regression models for categorical and limited dependent variables.* London: SAGE.

Rotnitzky, A. and Wypij, D. (1994) *A note on the Bias of Estimation with Missing Data.* Biometrika, **147**, *87-99.*

Wedderburn, R.W.M. (1974) *Quasi-likelihood functions, Generalized linear Models, and the Gauss-Newton method.* Biometrika, **61**, 439-447.

Zeger, S.L., Liang, K.Y. (1992) *An overview of methods for the analysis of longitudinal data.* Statistics in medicine, **11**, 1825-1839.

Zhao, L.P., Prentice, R.L. (1990) *Corrolated binary regression using a quadratic exponential model.* Biometrika, **77**, 642-648.